



e-ISSN:2582-7219



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 7, July 2024



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521



6381 907 438



6381 907 438



ijmrset@gmail.com



www.ijmrset.com



Job Post Verification using Machine Learning & Data Science

Kamireddygari Vinod Reddy, Prof. Rajesh N,

Student, Department of MCA, AMC Engineering College, Bannerghatta Road, Bangalore, India

Assistant Professor, Department of MCA, AMC Engineering College, Bannerghatta Road, Bangalore, India

ABSTRACT: The increasing prevalence of fake job postings on online job portals poses significant risks to job seekers, including financial loss, identity theft, and psychological distress. This study conducts study of various ML & Data science models to predict and detect fake job posts effectively. Utilizing a dataset of job postings from multiple online platforms, the research applies several algorithms, such as logistic regression, and networks of neural ..etc. Each technique's performance is evaluated based on key metrics such as Correct Classification Rate, positive predictive value, Hit Rate, and F-Measure. The evidence indicates that while all methods show some level of effectiveness, particular techniques excel over others in certain metrics. For instance, neural networks demonstrate high accuracy and recall but necessitate high computational capabilities & tuning. In contrast, logistic regression offers a more interpretable model with decent performance metrics. The study underscores the significance of choosing appropriate algorithms based on the specific requirements and constraints of the application. These insights aim to enhance the developing of robust systems for identifying and mitigating the risks associated with fake job postings, thereby protecting job seekers and maintaining the integrity of online job markets.

KEYWORDS: Fake Job Posts, Job post prediction, Data mining techniques, Prediction Models, Machine Learning, Data science, Feature Extraction, Text Classification, Comparative Analysis.

I. INTRODUCTION

The digital transformation has significantly revolutionized the job market, with numerous online platforms emerging as vital resources for job seekers and employers alike. These platforms, such as LinkedIn, Indeed, and Glassdoor, offer a vast array of job opportunities, making the job search process more accessible and efficient. However, this convenience also entails substantial risks, notably the proliferation of fake job postings. Fake job posts are deceptive listings created with malicious intent, such as extracting personal information, financial fraud, or spreading malware. As these fraudulent activities escalate, they pose severe threats to job seekers, including financial loss, identity theft, and emotional distress. Therefore, detecting and preventing fake job postings has emerged as a crucial field of study and application.

Data mining, an essential branch of artificial intelligence & machine learning, has shown considerable promise in addressing this issue. Data mining techniques involve analysing large datasets to discover patterns, anomalies, and relationships that can inform predictive models. These models used to classifies job postings as genuine or fraudulent, thereby mitigating the risks associated with fake job ads. This study aims to conduct a comparative analysis of various data mining methods to determine their effectiveness in predicting fake job posts.

Data Mining techniques offer a promising solution to this problem. By leveraging historical data and ML algorithms, with these techniques can analyse job postings and identify patterns indicative of fraudulent activity. Previous Research has shown the promise of various data mining methods in detecting fake job posts, but a comprehensive comparison of these techniques remains limited. This study aims to fill this gap by evaluating multiple data mining approaches, include support vector machines (SVM), decision tree, logistic regression, and neural networks, to identify Reliable forecasting approach fake job posts.

II. LITERATURE SURVEY

1. Literature Survey

The issue of fake job postings on online job portals had gained high attention from researchers, given the significant risks posed to job seekers and Comprehensive consistency across the job market. The proliferation of such fraudulent



activities necessitates the advancement of robust detection mechanisms. This literature survey reviews the existing body of research in the mining techniques, ML algorithms & data science to predict and detect fake job posts.

2. Early Studies on Fake Job Post Detection

Initial efforts in detecting fake job postings largely relied on manual methods and rule-based systems. These methods involved keyword filtering and heuristic approaches to identify suspicious job ads. Despite this, These methods often fell short because fraudsters constantly modify their strategies to evade basic detection filters. For instance, explored a heuristic approach that flagged job postings with common indicators of fraud, such as requests for personal information or up-front fees. While effective to some extent, this method lacked scalability and adaptability.

3. Challenges and Future Directions

Despite the achievements, there are persistent challenges in the detection of fake job postings. One major challenge is the evolving nature of fraudulent tactics, which requires continuous updating of detection models. Additionally, the imbalance in datasets, where genuine job postings far outnumber fake ones, poses difficulties for training effective models. Dealing with these challenges necessitates continuous exploration and study and the development of adaptive Models capable of acquiring knowledge from new data.

Future research directions include the exploration of ensemble methods involves amalgamating multiple models to facilitate improve predictive performance. The integration of NLP (natural language processing) techniques to analyse job descriptions and identify subtle cues of fraud is another promising avenue. Furthermore, the collaboration between academia and industry can facilitate the enhancement of more robust and scalable solutions to protect job seekers from fraudulent job postings.

III. EXISTING SYSTEM

The Identifying fraudulent job advertisements has evolved significantly with the advent of advanced data mining methods. The existing Systems crafted for this objective necessitate ongoing investigation leverage various MLalgorithms to analyse job postings and identify potential fraud. These systems generally follow a structured approach, comprising Data gathering, data preparation, and feature engineering, model training, and evaluation. Existing systems typically begin with the collection of job postings from various online platforms. These datasets often include job titles, descriptions, company names, locations, salary information, & other relevant attributes. A major difficulty at this point is acquiring labeled data, as it requires identifying which job postings are fake and which are genuine. This labeling Can be accomplished through manual means domain experts or through crowdsourcing.

Data preprocessing is essential to guarantee the quality and appropriateness of data for analysis. This phase involves cleaning the data by Eliminating redundancies, managing absent data entries, and standardizing text fields. Text pre-processing methods, such as tokenization, stemming, lemmatization, and the removal of stop words, are applied to job descriptions and other textual data. These steps help in reducing noise and enhancing the efficiency of the utilized features for model training.

IV. PROPOSED SYSTEM

To enhance the detection and prediction of fake job postings, The suggested system incorporates a multi-faceted An approach that harnesses sophisticated data techniques and addresses the limitations of existing systems. The suggested system aims to improve accuracy, adaptability, and computational efficiency while maintaining interpretability and scalability. The system structure comprises numerous key components: data acquisition, data scrubbing, variable selection, model training, evaluation, and real-time monitoring.

1. Data Collection and Preprocessing

The job verify begins with an extensive data collection phase, sourcing job postings from multiple online platforms, including popular job portals & social media websites. To get a robust dataset, the system utilizes web scraping techniques and APIs to gather a diverse range of job ads. The data is regularly refreshed to incorporate the most recent strategies and methodologies employed by scammers.

Preprocessing in the latest system includes advanced text normalization Methods like these tokenization, stemming, and lemmatization, coupled with advanced techniques for managing missing values and detecting outliers. The system



also employs entity recognition to extract meaningful entities, such as company names, locations, and job titles, improving the granularity and quality of the data.

V. METHODOLOGY

1. ML: Machine learning constitutes a segment of AI (artificial intelligence) focused on Creating algorithms that empower computers to learn from data & subsequently predict or make decisions based on data. ML algorithms allow systems to Autonomously acquire and enhance skills through experience, devoid of explicit programming. This capability is crucial In diverse applications, encompassing NLP and related methodologies, computer vision, healthcare diagnostics, and financial forecasting. Key techniques in ML include supervised learning (where algorithms learn from labeled data), unsupervised learning (where algorithms find patterns in unlabeled data), and reinforcement learning (where algorithms learn by interacting with an environment & receiving feedback). ML continues to advance rapidly, driven by increasing computational power, vast amount of available data, and innovative algorithmic developments, making it a cornerstone of modern technological innovation and automation.

Machine learning algorithms:

1. **Supervised Learning:** This type entails educating a model on a labeled dataset, meaning that each training example is paired with an output label. The model acquires knowledge learns to predict the output from the user information. Examples include classification & regression tasks.
2. **Unsupervised Learning:** In this type, The model undergoes training using datasets without labeled responses. The model tries to learn the underlying the organization of the data. Examples include clustering and association tasks.

2. Random Forest Classifier:

The Random Forest model: An ensemble method that combines multiple decision trees to improve prediction accuracy while decreasing the risk of overfitting. Each tree within the forest has trained autonomously on a subset of the dataset using randomized selection of features. During prediction, the ensemble aggregates the forecasts regarding individual trees to reach a final decision. This approach enhances robustness against noise and variance in the data, making Random Forests versatile For both categorization, grouping and regression tasks. They excel in handling large datasets with high-dimensional features, maintaining interpretability through feature importance rankings derived from tree-based computations. Random Forests are extensively utilized in different fields, including finance, healthcare, & ecology, where accurate and reliable predictions are crucial.

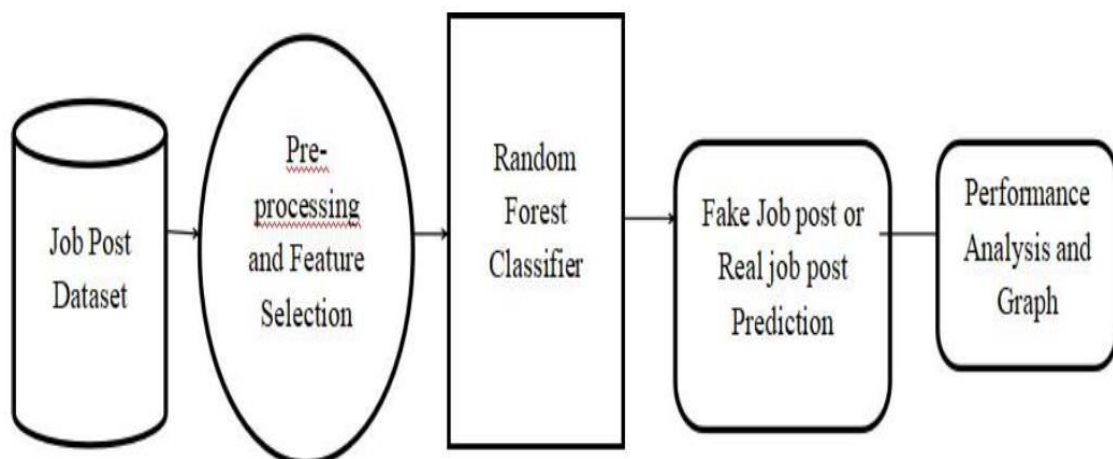


Fig 5.1 System model



VI. CONCLUSION

This study on predicting Counterfeit job listings Employing diverse data mining methodologies highlights the effectiveness and nuances of different models in tackling the prevalent issue of online job market fraud. Through comprehensive analysis and evaluation, it observed that each technique whether logistic regression, SVM, decision trees or neural networks offers distinct advantages in terms of precision and correctness, interpretability, and computational efficiency. Decision trees excel in transparency and simplicity, while SVMs provide robust classification in high-dimensional spaces. Logistic regression offers a balance between performance and interpretability, whereas neural networks demonstrate superior performance with complex data patterns. Through the comparison of these methods, this study provides significant and insightful findings for selecting appropriate techniques based on specific application requirements and data characteristics, thereby contributing toward the advancement of more reliable systems For identifying and alleviating counterfeit or fraudulent activities job postings in online platforms

REFERENCES

- [1] Vidros. S, Kolilias. C, Kambourakis. G, and Akoglu. L, “Automated identification of online recruitment scams: Traits, methodologies, and an open dataset”, *Future Internet* 2017, 9, 6; doi: 10.3390/fi9010006.
- [2] O. Nindyati and I. G. Bagus Baskara Nugraha, “Detecting Scam in Online Job Vacancy Using Behavioral Features Extraction,” in 2019 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia: IEEE, Nov. 2019, pp. 1–4. doi: 10.1109/ICISS48059.2019.8969842.
- [3] Ngan Luu-Thuy Nguyen¹, and Anh GiaTuan Nguyen, “Job Prediction From models based on neural networks to Applications”, *RIVF International Conference on Computing and Networking technologies (RIVF)*, 2020.
- [4] S. Dutta and S. K. Bandyopadhyay, “Detecting fraudulent job postings through a ML approach,” *InternationalJournal of Engineering Trends and Technology*, vol. 68, no. 4, pp. 48–53, Apr. 2020, doi: 10.14445/22315381/IJETT-V68I4P209S
- [5] F. Shibly, U. Sharma, and H. Naleer, “Performance Comparing Two-Class Boosted decision tree & Two-Class Decision Forest Algorithms for Predicting Fake Job Listings Postings,” *Annals of the Romanian Society for Cell Biology*, pp. 2462–2472, Apr. 2021.
- [6] R. S. Shishupal, Varsha, S. Mane, V. Singh, and D. Wasekar, “Efficient Utilizing Multinomial Naive Bayes for forecasting fake instances Job Profile,” *InternationalJournal of Advanced Research in could be paraphrased as Science, Communication and Technology*, pp. 286–291, May 2021, doi: 10.48175/IJARSCT-1241.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com